数据处理

数据规模

- 一共2193条用户数据，过滤掉出度（out_degree）加入度（in_degree）**小于等于**3的用户数据后还剩余2046个训练数据。

数据处理

数据主要包含3类，分别做以下处理：

- int value型数据（出度，入度，关注数，粉丝数，微博数, 注册时间）：
  - 采用函数$f(x) = log(x+1)$对数据进行初步处理，减少数据的极差。
  - 标准化处理（Z-Score）。
  - 缺失值处理：默认为0。
  - 其值直接作为特征向量中的一维。
- 离散型数据（领域、性别、是否认证）：
  - 这些离散型数据取值范围集合的基数较少，因此对这类数据进行独热编码。以性别为例，其编码为一个1*2的行向量，(1,0)代表性别为男、（0,1）则代表性别为女。
  - 缺失值（数据中仅有性别一列数据有缺失）处理：默认性别为男。
- 自然语言描述型数据（认证原因、简介、标签、地址）：
  - 使用NLP领域中的text2vec进行编码，对于每一个数据，其编码为一个1*96的向量。
  - 缺失值默认其编码向量为0。

特征向量化

将以上每一个数据都作为用户的一个特征，将这些特征进行拼接后得到每一个用户的特征向量。

特征矩阵shape：2045*397

降维与聚类

降维：

数据降维是指通过特征选择或者特征变换操作将数据从原始的D维空间投影到新的K维空间（K<<D），其基本作用有：

- 缓解维数灾难。即提高样本密度，以及使基于欧氏距离的算法重新生效。
- 数据预处理。对数据去冗余、降低信噪比。
- 方便可视化。

这里采用的是T-SNE算法。其基本原理是：

- 将数据点之间的相似度转化为`条件概率`，原始空间中数据点的相似度由`高斯联合分布`表示，嵌入空间中数据点的相似度由`t分布`表示。
- 通过原始空间和嵌入空间的联合概率分布的`KL散度损失函数`来评估嵌入效果的好坏。

最终将特征向量从**N*397映射到N*2**的低维向量空间中。

（无监督）聚类：

数据聚类是将数据按照一定标准分割为不同的类或者簇，使得同一个簇类的数据对象尽可能地相似，簇间数据对象有较好的区分度。

采用K-Means算法对数据进行聚类。其基本原理是计算K个簇中心和样本点的距离来不断地将样本点分到不同的簇中，同时更新K个簇中心并进行迭代计算。项目中K取3,4,5。

结果分析

聚类中心个数K=3

- 聚类结果：

- 特征分析

用户类别	人数	出度均值	入度均值	粉丝数均值	微博数均值	认证情况	涉及的领域情况	男女比
0	610	24.43	826.43	538.56W	31993	True:74.59% False:25.41%	1:76.93% 2:14.43% 3:4.75% 4:2.62% 5:1.80%	≈3:1
1	506	11.64	444.09	105.20W	11535	True:55.73% False:44.27%	1:0.92.29% 2:0.05.93% 3:0.01.78%	≈1:1
2	930	8.95	330.83	216.93W	22874	True:84.52% False:15.48%	1:85.05% 2:11.72% 3:2.69% 4:0.43% 5:0.11%	≈13:7

- 0类用户其收欢迎的程度（出度均值 入度均值 粉丝数均值）较高，活跃度较强（微博数均值），已认证与非认证比例为3:1，涉及的领域主要是领域1，有一部分是领域2，少部分为3,4,5。
    - 从认证原因上看，其主要是官方微博，以及一部分的人气很高的自媒体和个人博主。
    - 其地址信息大部分是缺失的（其他，或者认为是不定的）。
- 1类用户受欢迎程度最低，活跃度最低，已认证与非认证比例大约为5:4，涉及的领域主要是1，少部分2和3。
    - 其大部分是个人博主和自媒体。
- 2类用户受欢迎程度适中，活跃度适中，已认证与非认证比例大约为17:3，涉及领域主要是1，少部分2，极少部分3/4/5。
    - 主要是自媒体和个人博主。

聚类中心个数K=4

聚类结果：

特征分析：

用户类别	人数	出度均值	入度均值	粉丝数均值	微博数均值	认证情况	涉及的领域情况	男女比
0	301	8.93	505.8	178.19W	15092	True:100% False:0%	1:90.03% 2:7.64% 3:1.99% 4:0.33%	≈16:9
1	766	8.21	305.55	252.86W	23193	True:100% False:0%	1:83.16% 2:13.05% 3:3.26% 4:0.39% 5:1.3%	≈7:3
2	370	14.18	394.9	20.68W	13009	True:0.03% False:99.97%	1:94.86% 2:4.32% 3:0.81%	≈3:2
3	609	24.47	827.75	539.45W	32027	True74.71% False:25.29%	1:76.35% 2:14.45% 3:4.76% 4:2.63% 5:1.81%	≈3:1

- 0类节点：其受欢迎程度、活跃度适中，全部为认证的用户，主要涉及领域为1，少部分2，极少部分3和4。
    - 从认证原因上看，其大部分是个人博主。
- 1类节点:其受欢迎程度、活跃度中偏上，全部为认证的用户，主要涉及领域为1，一部分2，极少部分3和4，5。
    - 从认证原因上看，主要是个人博主和自媒体。
- 2类节点：其受欢迎程度、活跃度最低，几乎全部为未认证的用户，主要涉及领域是1，少部分2，极少部分3。
    - 未认证的用户缺少认证原因。
- 3类节点：类似于K=3中的0类节点。
    - 从认证原因上看，但是其自媒体和个人博主的数量更少了。
    - 从地址信息上看，其仍然是缺失的（不定的）。

聚类中心个数K=5

聚类结果：

特征分析：

用户类别	人数	出度均值	入度均值	粉丝数均值	微博数均值	认证情况	涉及的领域情况	男女比
0	314	32.06	783.79	798.52W	43650	True:90.13% False:9.81%	1:69.11% 2:17.2% 3:5.73% 4:4.46% 5:3.50	≈21:4
1	304	8.84	501.18	186.37W	15077	True:100% False:0%	1:90.13% 2:7.57% 3:1.97% 4:0.033%	≈16:9
2	763	8.23	306.60	249.89W	23232	True:100% False:0%	1:83.09% 2:13.11% 3:3.28% 4:0.39% 5:0.13%	≈7:3
3	295	16.38	874.55	263.69W	19655	True:58.31% False:41.69%	1:84.07% 2:11.53% 3:3.73% 4:0.68%	≈2:1
4	370	14.18	394.91	20.68W	13008	True:0.27% False:99.73%	1:94.86% 2:4.32% 3:0.81%	≈3:2

- 0类节点：受欢迎程度最高（且粉丝数很多），活跃度高，大部分为已认证用户，主要涉及领域1，一部分2，少部分3,4,5。
    - 从认证原因上看，其大部分为官方微博，极少部分的自媒体、博主。
    - 地址信息大部分缺失。
- 1类节点：受欢迎程度较高（粉丝数适中），活跃度较低。全部为已认证用户，主要涉及领域为1，少部分2，极少部分3和4。
    - 从认证原因上看，其主要是个人博主以及极少部分的自媒体。
- 2类节点：受欢迎程度适中，活跃低一般。全部为已认证用户，主要涉及领域为1，一部分2，极少部分的3,、4和5。
    - 从认证原因上看，其主要是个人博主和自媒体。
- 3类节点：受欢迎程度较高（但是粉丝数相当于0类来说较少），活跃度较低，认证用户与未认证用户比例相当。主要涉及领域是1，一部分2，极少部分的3和4。
    - 从认证原因上看，其主要是个人博主和自媒体。
    - 地址信息大部分缺失。
- 4类节点：其受欢迎程度、活跃度最低，几乎全部为未认证的用户，主要涉及领域是1，少部分2，极少部分3。
    - 未认证的用户缺少认证原因。